



# CS 106S Week 7

Trust and Safety

[cs106s.stanford.edu](https://cs106s.stanford.edu), Autumn 2024

**Is the Internet *safe*?**

# Gamergate

THE TACTICS

## When the Internet Chases You From Your Home

By Sarah Jeong

Ms. Jeong is a member of the editorial board.

AUG. 15, 2019



**Drac** @CountDraculaNES **Spooky** 9h  
I just put a dead squirrel in Zoe  
Quinn's mailbox lmao [#GamerGate](#)

[Details](#)



**Network Jesus**  
@networkjesus



[Follow](#)

[#ZoeQuinn](#) is a lying, manipulative [redacted] and  
should be doxxed and destroyed. [redacted] any  
and all people that are covering her  
worthless [redacted].

[Reply](#) [Retweet](#) [Favorite](#)  
7:56 AM · 19 Aug 2014



**OpBlitzkrieg**  
@OpBlitzfag

@[TheQuinnspiracy](#) it'll be even weirder once my .55m  
reaches your cranium, talk to you soon, zoe-poo.



**Sanic**  
@Sanic\_Boom

@[TheQuinnspiracy](#) lmfao you know you're going to  
get [redacted] jumped when you go there right

9:17 PM · 25 Aug 14



**CabronMasterBait** @ElMasterBait · 22s  
@[TheQuinnspiracy](#) Im not only a pedophile, ive [redacted] countless teens, this  
zoe [redacted] is my next victim, im coming [redacted]

8:58 PM · 19 Aug 2014 · [Details](#)

[Hide conversation](#)

[Reply](#) [Retweet](#) [Favorite](#) [More](#)

# Disney forces explicit Club Penguin clones offline

15 May 2020

Share  Save 

**Joe Tidy**

Cyber-security reporter

Server owners accused one another of hacking and harassment. One said it was a toxic community, "like Game of Thrones with penguins".

The Club Penguin Online volunteer claims he was encouraged to carry out attacks on rival servers when he was a minor.

- content filters designed to remove offensive language had been disabled on several servers, allowing swear words, homophobic slurs, anti-Semitism and racist messages to be posted publicly
- players were engaging in "penguin e-sex", sending and receiving explicit messages

**not good.**



# What's Trust and Safety?

- 1 The study of how people abuse the internet to cause harm.
- 2 Often using products the way they are designed to work.
- 3 Crosses between specialties. Requires understanding of society and humanity.
- 4 Is dynamic and unpredictable

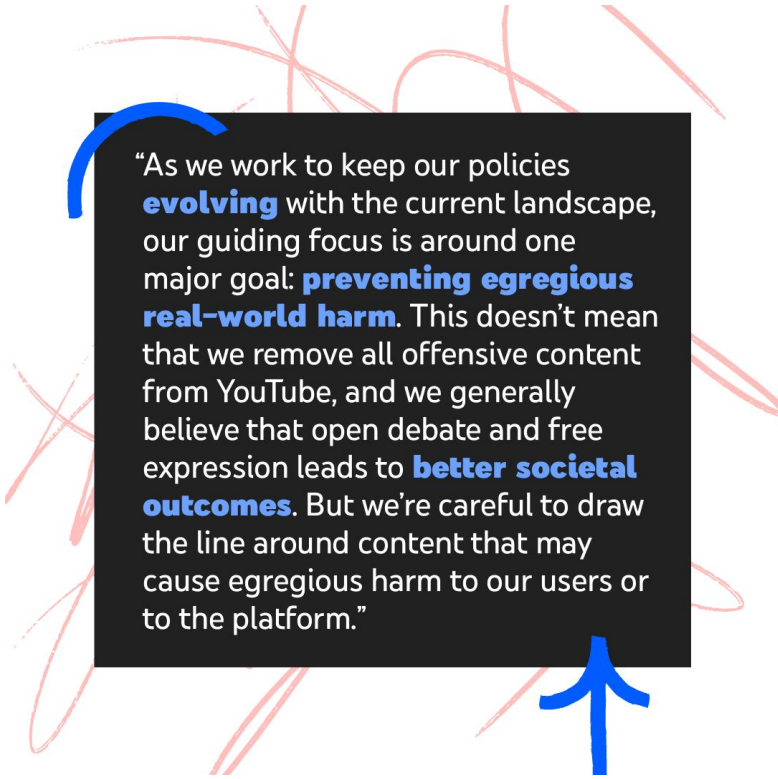


# Case Studies


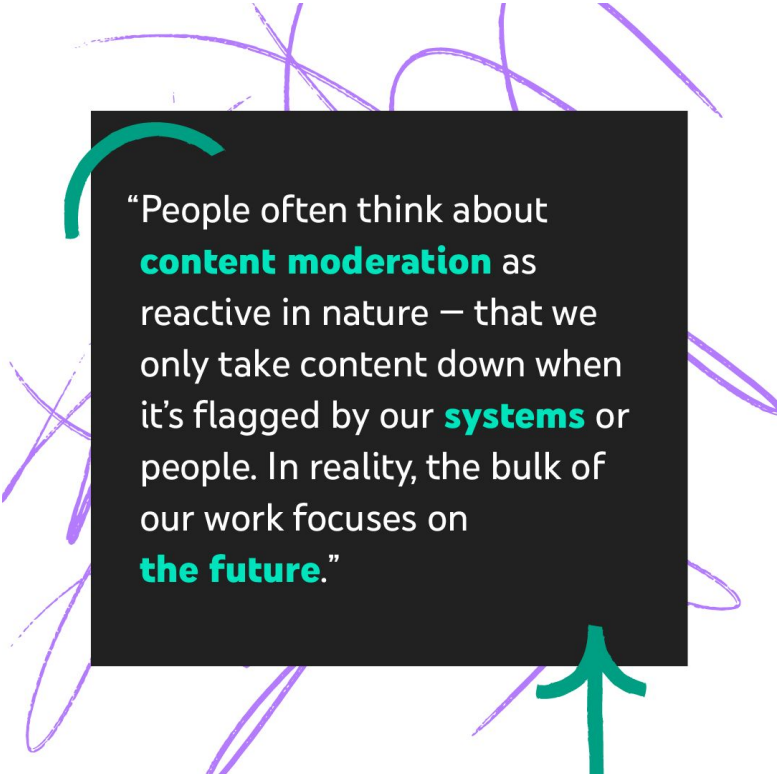
# Case Study #1

## Situation


People are uploading highly graphic content to YouTube.

A series of red, hand-drawn scribbles and loops surrounding the left text box.

“As we work to keep our policies **evolving** with the current landscape, our guiding focus is around one major goal: **preventing egregious real-world harm**. This doesn’t mean that we remove all offensive content from YouTube, and we generally believe that open debate and free expression leads to **better societal outcomes**. But we’re careful to draw the line around content that may cause egregious harm to our users or to the platform.”

A blue arrow pointing upwards towards the bottom of the left text box.A series of purple, hand-drawn scribbles and loops surrounding the right text box.

“People often think about **content moderation** as reactive in nature — that we only take content down when it’s flagged by our **systems** or people. In reality, the bulk of our work focuses on **the future**.”

A green arrow pointing upwards towards the bottom of the right text box.

<https://blog.youtube/inside-youtube/policy-development-at-youtube/>



# Case Study #2

## Situation

ISIS websites are showing up on Google Search results.

Jigsaw, the Google-owned tech incubator and think tank---until recently known as Google Ideas---has been working over the past year to develop a new program it hopes can use a combination of Google's search advertising algorithms and YouTube's video platform to target aspiring ISIS recruits and ultimately dissuade them from joining the group's cult of apocalyptic violence. The program, which Jigsaw calls the Redirect Method and plans to launch in a new phase this month, places advertising alongside results for any keywords and phrases that Jigsaw has determined people attracted to ISIS commonly search for. Those ads link to Arabic- and English-language YouTube channels that pull together preexisting videos Jigsaw believes can effectively undo ISIS's brainwashing---clips like testimonials from former extremists, imams denouncing ISIS's corruption of Islam, and surreptitiously filmed clips inside the group's dysfunctional caliphate in Northern Syria and Iraq.

<https://www.wired.com/2016/09/googles-clever-plan-stop-aspiring-isis-recruits/>



## For some searches, prioritize website Quality (i.e. trustworthiness or credibility) over Relevance

Meaning

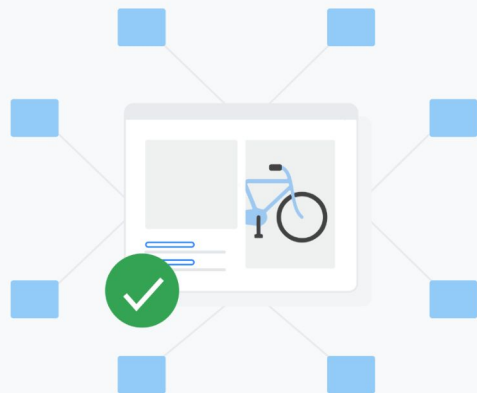
Relevance

Quality

Usability

Context

### Quality of content



After identifying relevant content, our systems aim to prioritize those that seem most helpful. To do this, they identify signals that can help determine which content demonstrates expertise, authoritativeness, and trustworthiness.

For example, one of the factors used to determine quality is understanding if other prominent websites link or refer to the content. This is generally a good sign that the information is trustworthy. Aggregated feedback from our Search quality evaluation process helps to refine how our systems discern the quality of information.

<https://www.google.com/intl/en/search/howsearchworks/how-search-works/ranking-results/#quality>

**Yes I know, jarring tone shift – but make sure to have mochi before you leave class today!**

